Fine-Tuning a LLM with a Brazilian Law Dataset

João Pedro Monteiro Volpi
CentraleSupélec
joao-pedro.monteiro-volpi@student-cs.fr

Lucas Vitoriano de Queiroz Lira
CentraleSupélec
lucasvitoriano.queirozlira@student-cs.fr

Abstract

This paper presents a project aimed at refining an NLP model for the Brazilian legal system. Leveraging the Google Gemma NLP model and Microsoft Phi-2 model, pre-trained on diverse corpora, we adapt it to process Brazilian Portuguese legal texts using the Iudicium Textum Dataset (1) from the Brazilian Supreme Court. Our goal is to create a tool for legal research, case analysis, and judicial decision prediction. We fine-tune the model on the legal dataset, incorporating domain-specific nuances, and evaluate its performance against baselines. This project contributes to NLP advancements and addresses the need for tailored linguistic solutions in Brazilian law.

1. Introduction

The ever-growing volume of legal documents poses significant challenges for legal professionals and researchers navigating the complex landscape of the Brazilian legal system. Traditional methods of legal research can be time-consuming and resource-intensive, highlighting the need for innovative solutions.

Existing NLP models often struggle with the nuances of legal language in Brazil, creating a research gap. This project addresses this gap by focusing on the development of an NLP model specifically designed for the Brazilian legal system.

Through this work, we aim to refine existing NLP models like Google Gemma and Microsoft PHI-2, leveraging the Iudicium Textum Dataset. This fine-tuned model will be capable of assisting in legal research, analyzing cases, and understanding judicial decisions within the Brazilian context. This tool has the potential to significantly enhance the efficiency and effectiveness of legal research for professionals and researchers, ultimately contributing to the advancement of NLP technologies within the Brazilian legal domain.

2. State-of-the-Art

The landscape of Large Language Models (LLMs) has seen rapid advancements, with significant contributions from major research institutions and tech companies in LLM of all sizes. Some independent researches have fine-tuned these primarily english models in the portuguese language.

2.1. Sabiá

Sabiá (2) represents a landmark achievement by Maritaca AI, challenging the prevailing "one-size-fits-all" model approach with a focus on monolingual pre-training for Portuguese. This model, derived from further training on the GPT-J and LLaMA architectures, leverages Portuguese texts to achieve significant performance improvements. The Sabiá-65B model, in particular, demonstrates comparable performance to GPT-3.5-turbo in some metrics, showcasing the power of language-specific pre-training in capturing linguistic nuances and domain knowledge inherent to Portuguese. The model's efficacy is further evidenced by its performance on Poeta, a suite of 14 Portuguese datasets, where it outperforms English-centric and multilingual counterparts. Sabiá's approach underscores the importance of domain-specific knowledge, revealing that the majority of benefits stem from monolingual pre-training rather than linguistic structure capture alone.

2.2. Cabrita

Cabrita (3) introduces a novel methodology aimed at enhancing performance and efficient tokenization at a manageable cost. The approach, applicable to any transformer-like architecture, involves continuous pre-training exclusively on Portuguese text using a model named OpenLLaMA, resulting in the creation of OpenCabrita 3B. This model addresses performance concerns and introduces a new tokenizer that significantly reduces the number of tokens required to represent text, offering a solution to one of the primary challenges faced by LLMs in specific domains or languages. In their assessment, for few-shot learning tasks, they achieved similar results with this 3B model compared to a traditional continuous pre-training approach as well as to 7B models English pre-trained models.

2.3. Bode

Bode (4), building on the foundation laid by Cabrita, extends the fine-tuning approach to incorporate the advancements of the LLaMA 2 model by Meta AI. This model aims to enhance Portuguese prompt response applications through a comprehensive dataset encompassing a broad range of Portuguese instructions. The fine-tuning process utilizes Low-Rank Adaptation, optimizing the model for interactive, reliable communication, and accurate response generation.

3. Background

3.1. Dataset

The Iudicium Textum Dataset is the base of our project, consisting of a comprehensive corpus of legal decisions (”acórdãos”) from the Brazilian Supreme Federal Court (STF). This dataset was designed to overcome the scarcity of Portuguese language resources for robust NLP model training. With over 40,000 decisions, categorized by subject matter and authorship, the dataset makes it possible to advance in text classification, topic modeling, and other NLP applications in the legal domain. The dataset has a lot of information about Brazilian legal language, making it a good foundation for training our NLP model.

The ”acórdãos” are formal decisions issued by courts, particularly by higher courts such as the Supreme Federal Court (STF) in Brazil. These documents are the result of collective judgments made by a panel of judges serve as official records that detail the facts of the case, the legal reasoning, and the final decision. Acórdãos are crucial in the legal system for several reasons:

- **Legal Precedent:** They serve as precedents for future cases, guiding lower courts and future panels in similar legal matters.
- **Legal Research:** Researchers, lawyers, and legal scholars use acórdãos to study legal reasoning, understand trends in judicial decisions, and develop legal theories.
- **Public Record:** They provide transparency and accountability in the judiciary, allowing the public to access the reasoning behind legal decision

Another studied dataset, the VICTOR dataset (5) is composed of 45 thousand Extraordinary Appeals *(Recursos Extraordinários)* from the STF. Each suit in turn contains several different documents, ranging from the appeal itself to certificates and rulings, totaling 692,966 documents comprising 4.6 million pages. The suits were manually annotated by experts from the Court staff with labels for the document classes.

Although this dataset was a interesting resources for our legal fine-tuning, we did not use it due to computing power constraints.

3.2. Google Gemma

Google DeepMind’s Gemma represents a significant advancement in the field of neural network-based language models, encompassing a series of models with 2 billion (2B) and 7 billion (7B) parameters. This paper utilizes the 2B variant, chosen due to the computational limitations of our environment.

Context Length: With an impressive context length of 8192 tokens, Gemma can comprehend and generate lengthy passages of text, making it adept at handling extensive narratives or detailed documentations.

Dataset Size: Trained on a robust dataset featuring 2 terabytes (T) of tokens primarily sourced from English web documents, mathematical texts, and coding materials. This extensive dataset ensures Gemma’s ability of interpreting and producing a wide array of text types.

Environmental Impact: The pretraining phase of Gemma has been assessed to emit approximately 131 metric tons of CO2 equivalents (tCO_2 eq), calculated from the energy consumption data of Google’s TPU data centers, including the overheads for maintaining the data center infrastructure.

3.3. Microsoft Phi-2

The Phi-2 model(6), developed by Microsoft Research, stands as a transformative addition to the expansive array of language models within the Phi series, equipped with 2.7 billion parameters. This initiative is aimed at rivaling the capabilities of considerably larger models while emphasizing efficiency and scalability.

Context Length: Phi-2 is engineered to process text sequences up to 2048 tokens in length, facilitating its ability to tackle detailed textual analyses and generation tasks that require understanding of extended contexts.

Dataset Size: The model’s training regimen incorporates a composite dataset of 250 billion tokens, merging synthetic NLP data and meticulously filtered web content. This dataset amalgamation serves to enhance the model’s linguistic versatility and application across a multitude of settings.

Environmental Impact: While the Phi-2 model documentation details various aspects of its development, including training data, architecture, and limitations, information regarding its environmental impact is currently not publicly available.

3.4. Low-Rank Adaptation

Low-Rank Adaptation (LoRA)(7) is a technique designed to facilitate the adaptation of LLMs extensively pre-trained on a vast corpus of textual data to novel and unexplored tasks without incurring any inference latency or compromising the length of input sequences.

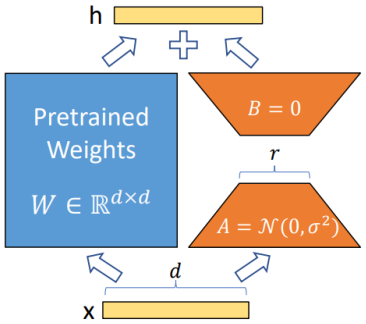

Figure 1. LORA finetuning addresses performance challenges through the use of a low-rank matrix

When confronted with a new downstream task, LoRA inserts a compact low-rank matrix into the pre-trained weight matrices of the language model. Then, these newly added parameters are trained with the specific downstream task and serve as the pivotal element for adapting the pre-trained weights to the unique requirements of the new task at hand. Compared to the standard fine-tuning procedure, LoRA significantly reduces the trainable parameters required 1 for adaptation to downstream tasks while maintaining a high-quality model.

Trainable Params	All Params	Trainable %
156,893,184	2,787,548,160	5.891

Table 1. Lora Effect on the Phi-2B model

4. Our Approach

This section explores the open-source tools utilized and the techniques for developing an finetuned Large Language Models (LLMs). We employed the Hugging Face Transformers library(8) to access and fine-tune pre-trained models, significantly reducing the time and resources required for custom model training. Additionally, we leveraged PyTorch 2.2 for its deep learning framework and CUDA 12.2 for GPU acceleration, optimizing the training process. This combination of tools, along with a secure GPU cloud platform, enabled us to surpass the constraint of the available 30h weeks use of free GPU providers. The following sections will detail our methodology, experimental setup, and results.

4.1. Fine-tuning

In the fine-tuning process, certain parameters were set to adapt the model to specific tasks and to manage hardware constraints.

4.1.1 Gradient Accumulation Steps

After processing each mini-batch, instead of updating the model weights immediately, the gradients are accumulated over several steps. Once the specified number of steps is reached, the accumulated gradients are used to update the model weights. Simulating the effects of training with large batch sizes, which can lead to more stable and reliable training convergence.

4.1.2 Optimizer (AdamW_Torch)

AdamW(9) is an optimizer that combines the best features of two other popular optimizers: AdaGrad and RMSProp, and introduces weight decay for regularization, a technique to prevent overfitting by penalizing large weights. The inclusion of weight decay helps in separating the weight's regularization from the optimization steps, leading to more effective fine-tuning of models

4.1.3 LR Scheduler (Cosine) with Warmup

A learning rate scheduler adjusts the learning rate during training, and the cosine scheduler with warmup specifically starts with a low learning rate, gradually increases it during the warmup phase to a maximum value, and then decreases it following a cosine curve. This approach helps the model to start learning slowly, preventing premature convergence to suboptimal minima, and then leverages the cyclical nature of the cosine function to navigate the loss landscape more effectively.

4.1.4 Flash Attention 2 ("flash_attention_2")

Flash Attention (10) is an optimized attention mechanism designed to reduce the computational cost and memory usage associated with traditional attention1 calculations in Transformer models. It works by efficiently batching the computation of attention scores and leveraging hardware accelerations. The "flash_attention_2" implementation provides an even more optimized version, further reducing computational overhead and enabling faster training and inference times.

$$\text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{1}$$

4.1.5 Saving Strategy (save_strategy)

Is a technique used to save the trained model only if the performance on a validation set improves for a specified number of epochs.

4.2. Adapting to Hardware Constraints

To address hardware limitations during the fine-tuning process, we followed certain strategies:

4.2.1 R (Rank)

The rank in Low-Rank Adaptation (LoRA) impacts the model's computational complexity changing directly the size of the weight matrix that will be optimized as showed in the image 1. A lower rank,such as $r = 64$, simplifies the update mechanism within the model, making it computationally less expensive while still maintaining a high level of performance. This parameter is essential for deploying large models on devices with limited computational resources, enabling the use of advanced models in a wider range of applications.

4.2.2 Lora_Alpha

The $lora\_alpha$ parameter scales the magnitude of updates in LoRA, striking a balance between the aggressiveness of the updates and the model's stability. Setting $lora\_alpha = 64$ helps in fine-tuning the model's learning process, ensuring that the updates are substantial enough to improve the model without causing instability in its predictions.

4.2.3 Context Length

The context length parameter, set to 512 in this configuration, directly influences the model's ability to process and generate sequences of text. A longer context length allows the model to consider a wider window of text for its predictions, potentially improving its understanding of context and its ability to generate coherent and relevant outputs. However, increasing the context length requires more memory and computational power.

4.2.4 4-bit Quantization

"Load in 4 bit" or 4-bit quantization significantly reduces memory consumption by compressing the model weights without a considerable loss in performance. This approach is crucial for deploying large models on devices with limited memory, making it possible to leverage state-of-the-art models in constrained environments. By using 4-bit quantization, models become more accessible for use in real-world applications where hardware limitations are a concern.

4.3. Training

Throughout the training period, the computational resources included an **RTX4090** GPU, boasting 24GB of VRAM, which facilitated the execution of the training regimen. This phase spanned roughly 5 hours, throughout which the model engaged in processing an aggregate of 10,240 tokens. Additionally, the hardware configuration was complemented by a 32 Core processor and 103GB of available RAM, optimizing the overall training environment for efficiency and speed.

For the training procedure we used the following prompt:

> *This is a text about a court decision made by the Brazilian Supreme Court [text from dataset].*
>
> *Question: What laws, articles, and legal instruments are mentioned, used, and applied in this court decision?*
>
> *Answer: [model output]*

5. Experiments / Results

In this section, we present the empirical results obtained from the fine-tuning process of the Gemma and Phi-2 models, within the context of our study, and also evaluate the finetuned results using the GPT 3.5 Turbo.

5.1. Comparing Finetune

To assess the performance impact of hyperparameter tuning on model fine-tuning, we embarked on a systematic examination, varying two critical parameters: learning rate and LoRA dropout. For this purpose, we conducted experiments across a spectrum of settings for these parameters, specifically choosing learning rate values of 2^{-4} and 2^{-5}, and dropout rates of 0.1 and 0.2. This approach allowed us to directly observe the effects of these hyperparameters on the model's learning efficiency and generalization capability over the course of fine-tuning.

Given the computational intensity and performance considerations of training numerous models, our exploration was pragmatically limited. We meticulously selected a range of models that embody a broad overview of the configurations tested. Each model underwent fine-tuning over a span of 21 epochs, a duration determined based on the observation that optimal performance generally materializes within the initial epochs.

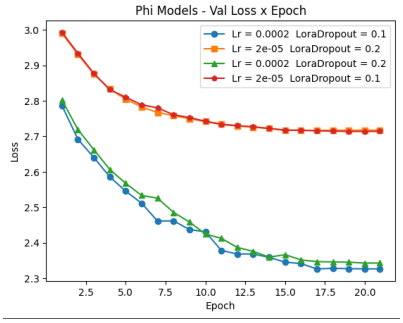

Figure 2. Phi-2 Model Validation Loss over epoch

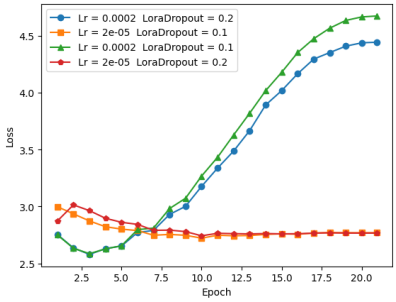

Figure 3. Gemma-2B Model Validation Loss over epoch

5.2. Evaluation

For the assessment phase, we crafted 150 questions pertaining to Brazilian Law in Portuguese by utilizing OpenAI's GPT 3.5. In sequence, after training four models each with different parameters for Phi-2, and another four for Gemma, we identified the finest-tuned model from both sets, characterized by the least loss. To compare these two models, we employed the following methodology:

Utilizing OpenAI's GPT 3.5 Turbo API, we presented the model with the subsequent prompt:

> *I'm evaluating two language models to decide which is best, and I need your help. I will show you the prompt I sent and the model's response for both models, and I need you to respond to me by telling me which model had the best response.*
>
> *Model 1: [response1]*
>
> *Model 2: [response2]*
>
> *Which model had the best response? Answer with just one digit, indicating the model number (1 or 2).*

In this manner, we derived a score for each model based on an evaluation conducted by a significantly larger model.

The evaluation outcomes are summarized in the table below. The numbers represent the number of times the model's answers were preferred in relation with with the other model's answers (the row sums up to 150 questions).

Gemma (FT)	Phi (FT)
132	18

Table 2. Response preferences between fine-tuned Gemma and fine-tuned Phi model

Subsequently, we subjected the "winning" model, Gemma, against its base model (without fine-tuning) to discern the efficacy of the fine-tuning process.

The comparative results are presented below:

Gemma	Gemma (FT)
64	86

Table 3. Response preferences between base Gemma and our fine-tuned Gemma model

The findings demonstrate that while the fine-tuned model exhibited superior performance, even slight variations in the initial prompt substantially influenced the selection of the best response in GPT 3.5. In contrast, when comparing Gemma and Phi, Gemma consistently favored responses to a significant degree. These observations underscore the necessity for a more precise metric capable of accurately assessing response quality.

One potential avenue for improvement is leveraging Human Feedback. However, this methodology is fraught with challenges, notably its time-intensive nature and limited effectiveness.

5.3. Prompt Example

In this section, we present an example of how we asked a question during testing. Despite none of the answers being perfect, we'll observe that the second one is closer to being correct and makes more sense.

> *You are an expert in Brazilian law. Answer the following question in Portuguese using your knowledge of Brazilian law, the constitution and its articles, the Federal Supreme Court, etc.*
>
> *Question: What are the requirements for someone to be found guilty in a criminal case in Brazil?*

Base Gemma:

> *"Article 11 of the 1988 Federal Constitution establishes that Brazilian criminal law is based on law, which is the expression of legislative power. Article 12 of the Federal Constitution of 1988 establishes that the executive power is the executive of the State, which is the executive power of Brazil. Article 13 of the Constitution Federal Law of 1988 establishes that the judiciary is the judiciary of the State, which is the judiciary of Brazil. Article 14 of the Constitution Federal Law of 1988 establishes that the legislative power is the legislative power of the State, which is the legislative power of Brazil. Article 15 of the Constitution Federal Law of 1988 establishes that the executive power is the executive power of the State, which is the executive power of the"*

Fine-Tuned Gemma:

"To be convicted in criminal proceedings in the Court of Justice of the State of São Paulo, any character must demonstrate the existence of any deficiency in any constitutional demand or constitutional rule of criminal procedure law. The existence of a deficiency in the constitutional demand for the right to criminal proceedings depends on any constitutional demand for the right to criminal proceedings being recognized by the Court of Justice of the State of São Paulo. For existence"

6. Limitations

One of the principal challenges faced in our study revolved around the extensive computational resources necessitated for the efficient fine-tuning of the language model, as well as for leveraging the full extent of the dataset at our disposal. Given that the VictorDataset(5) comprises an immense volume of data, tallying up to 170 million tokens, the task of training with the entirety of the available information proved to be computationally prohibitive. The sheer scale of the dataset demanded not only significant processing power but also considerable memory and storage capacity, which were beyond the scope of our available infrastructure.

7. Conclusion

In conclusion, our exploration into fine-tuning a Large Language Model with a Brazilian Law Dataset has revealed insights into the adaptability and performance of NLP models within the legal domain. By varying learning rates and LoRA dropout rates, we evaluated the impact of these hyperparameters on model efficiency over 21 epochs. Our findings show the importance of hyperparameter optimization in enhancing model performance for specific applications. Through comparative evaluation, we have demonstrated the potential of fine-tuning techniques to improve the utility of NLP models for legal research in the Brazilian context, although further work must still be done. This work hopes to contribute to the advancement of NLP in legal systems and to pave the way for future research focused on the application of language models in specialized domains.

References

[1] A Willian Sousa and Marcos Didonet Del Fabro. Iudicium textum Dataset: Uma Base de Textos jurídicos para NLP [iudicium textum dataset: A legal text base for nlp]. pages 427–434, 2020. 1

[2] Ramon Pires, Hugo Abonizio, Thales Sales Almeida, and Rodrigo Nogueira. Sabiá a Portuguese Large Language Models. volume 14197, pages 226–240. 2023. arXiv:2304.07880 [cs]. 1

[3] Celio Larcher, Marcos Piau, Paulo Finardi, Pedro Gengo, Piero Esposito, and Vinicius CaridÃ¡. Cabrita: closing the gap for foreign languages, August 2023. arXiv:2308.11878 [cs]. 1

[4] Gabriel Lino Garcia, Pedro Henrique Paiola, Luis Henrique Morelli, Giovani Candido, Arnaldo Cândido Júnior, Danilo Samuel Jodas, Luis C. S. Afonso, Ivan Rizzo Guilherme, Bruno Elias Penteado, and João Paulo Papa. Introducing Bode: A Fine-Tuned Large Language Model for Portuguese Prompt-Based Task, January 2024. arXiv:2401.02909 [cs]. 2

[5] Pedro Henrique Luz de Araujo, Teófilo Emídio de Campos, Fabricio Ataides Braz, and Nilton Correia da Silva. VICTOR: a dataset for Brazilian legal documents classification. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1449–1458, Marseille, France, May 2020. European Language Resources Association. 2, 6

[6] Model catalog - Azure AI Studio. 2

[7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models, October 2021. arXiv:2106.09685 [cs]. 2

[8] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing, October 2020. 3

[9] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, January 2019. arXiv:1711.05101 [cs, math]. 3

[10] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness, June 2022. arXiv:2205.14135 [cs]. 3